applied sciences

MDPI

Article

The Domain Mismatch Problem in the Broadcast Speaker Attribution Task

Ignacio Viñals *,†, Alfonso Ortega *,†, Antonio Miguel *,† and Eduardo Lleida *,†

ViVoLab, Aragón Institute for Engineering Research (I3A), University of Zaragoza, 50018 Zaragoza, Spain

* Correspondence: ivinalsb@unizar.es (I.V.); ortega@unizar.es (A.O.); amiguel@unizar.es (A.M.); lleida@unizar.es (E.L.)

† These authors contributed equally to this work.

Abstract: The demand of high-quality metadata for the available multimedia content requires the development of new techniques able to correctly identify more and more information, including the speaker information. The task known as speaker attribution aims at identifying all or part of the speakers in the audio under analysis. In this work, we carry out a study of the speaker attribution problem in the broadcast domain. Through our experiments, we illustrate the positive impact of diarization on the final performance. Additionally, we show the influence of the variability present in broadcast data, depicting the broadcast domain as a collection of subdomains with particular characteristics. Taking these two factors into account, we also propose alternative approximations robust against domain mismatch. These approximations include a semisupervised alternative as well as a totally unsupervised new hybrid solution fusing diarization and speaker assignment. Thanks to these two approximations, our performance is boosted around a relative 50%. The analysis has been carried out using the corpus for the Albayzín 2020 challenge, a diarization and speaker attribution evaluation working with broadcast data. These data, provided by Radio Televisión Española (RTVE), the Spanish public Radio and TV Corporation, include multiple shows and genres to analyze the impact of new speech technologies in real-world scenarios.

Keywords: speaker attribution; diarization; multi-domain; domain mismatch

Citation: Viñals, I.; Ortega, A.; Miguel, A.; Lleida, E. The Domain Mismatch Problem in the Broadcast Speaker Attribution Task. *Appl. Sci.* **2021**, *11*, 8521. https://doi.org/10.3390/app11188521

Academic Editor: José A. González-López

Received: 4 August 2021
Accepted: 9 September 2021
Published: 14 September 2021

1. Introduction

In recent years, the importance of extensive high-quality annotations for multimedia data has been highlighted. Regarding audio, this information may describe the inner details of the data (who talks in an audio, what is said, etc.) or link multiple pieces according to external characteristics (e.g., someone's music preferences). Thus, the net worth of a certain set of multimedia data now heavily depends on the quality of its labels.

Unfortunately, the inference of more and more elaborated labels is not free of charge. In the past, metadata generation was hand-made, restricted to few types and expensive in terms of money, time and effort. These days, the variety of requested side information has largely increased and must be obtained for a much larger amount of multimedia content. Hence, large-scale human annotation tends not to be viable. Fortunately, thanks to the evolution of artificial intelligence, computer-based or computer-assisted solutions can be considered instead.

One of the most traditional inferred audio metadata is speaker information. However, not all the speaker metadata is equally descriptive. Speaker attribution is the job responsible for inferring whether any speaker from a set of enrolled speakers contributes to the audio under analysis. Furthermore, if these speakers are present, the estimation of timestamps when these contributions occur is also requested. For this purpose, the speaker enrollment should be according to a small portion of audio per person of interest. Thus, speaker attribution can be interpreted as the fusion of two well-known tasks: speaker verification and diarization.

The study of speaker attribution has been largely studied, as shown in the bibliography. Usually, it has been considered an evolution of diarization. Thus, very often, it follows a similar Bottom-Up strategy: First, the input audio is divided into segments under the assumption of a single speaker per audio fragment. Then, each segment is transformed into a speaker representation and assigned to its corresponding cluster according to its speaker. Within the bibliography, speaker attribution has analyzed multiple sorts of speaker representations: Joint Factor Analysis (JFA) [1], i-vectors [2], PLDA [3], or DNN embeddings, such as x-vectors [4]. Regarding systems, some alternatives rely on Agglomerative Clustering [5–7] taking into account different metrics (cosine distance, Kullback–Leibler divergence, Cross Likelihood Ratio (CLR), etc). Other contributions [8,9] exploit Information Theory concepts, such as Mutual Information, to make decisions. The assignment of clusters by means of a speaker recognition paradigm has also been proposed in [10]. Finally, ref. [11] proposes graph-based semi-supervised learning approximation to speaker attribution. Most of these systems consider well-known techniques and state-of-the-art approximations. However, they usually suffer from a similar limitation: lack of robustness. These systems rely on a threshold under the assumption of similar conditions in both development and evaluation scenarios.

The motivation for this article is the analysis of speaker attribution in broadcast data. The broadcast domain and its archive services are keen on many types of automatic annotation, including speaker attribution. This interest is due to the increase of produced content experiments within recent years as well as the high complexity of the domain nature. This content, understood as a collection of shows and genres with particular characteristics, provides a challenging domain that requires techniques capable of dealing with such variability. In fact, this complexity motivates the division of our main objective, the analysis of speaker attribution, into three partial goals. These partial goals are:

- To study the influence of diarization on the performance of speaker attribution systems;
- To analyze the impact of domain mismatch between models and data;
- To propose robust approximations that mitigate the domain mismatch between models and data under analysis.

The audio used in this study belongs to the latest of the ongoing series of Albayzín evaluations [12]. These evaluations seek the evolution of speech technologies, such as Automatic Speech Recognition (ASR), diarization and speaker attribution, with special emphasis on the broadcast domain. For this purpose, the whole corpus, gathered along the multiple editions, consists of audio from real broadcast content from radio stations and TV channels. Regarding the 2020 edition, the data are released by Radio Televisión Española (RTVE), the Spanish public Radio and Television Corporation (RTVE collaboration through http://catedrartve.unizar.es/, accessed on 13 September 2021).

This article is organized as follows: A study of the speaker attribution problem is carried out in Section 2. The experimental scenario is described in Section 3. The studied systems are explained in detail in Section 4. Section 5 is dedicated to the results of the experiments carried out in this article. Finally, our conclusions are collected in Section 6.

2. The Speaker Attribution Problem

The speaker attribution problem is a complex task focused on inferring detailed speaker information about any input audio. As illustrated in Figure 1, speaker attribution should estimate whether a set of enrolled people contribute to a given audio and, if so, when they talk. These decisions are made according to small portions of audio from the speakers of interest. Whenever all speakers in the audio are reassured to be enrolled in the system, we are dealing with a closed-set scenario; otherwise, it is an open-set scenario. Some contributions in the field of speaker attribution are [13], which focuses on the identity assignment and includes purity quality metrics, [14], which combines the diarization task with a speaker-based identity assignment, and [15], which speeds up the diarization and the identity assignment process by means of low-resource techniques.

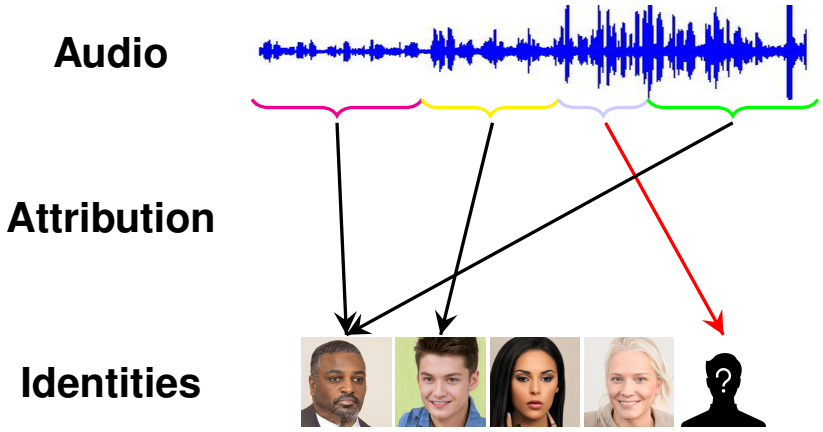

Figure 1. Concept diagram of speaker attribution. For the given audio, we assign the portion of speech generated by each enrolled speaker. Additionally, we must also detect the audio belonging to non-enrolled speakers (red arrow).

State-of-the-art speaker attribution has been built by collecting techniques to identify the audio from a single speaker in a recording and strategies to assign this audio to the corresponding enrolled speaker. This description also fits other tasks, such as speaker linking and longitudinal diarization, which require similar techniques and only differ in the description of enrollment audios. A general solution for all these tasks follows a diagram block, as shown in Figure 2.

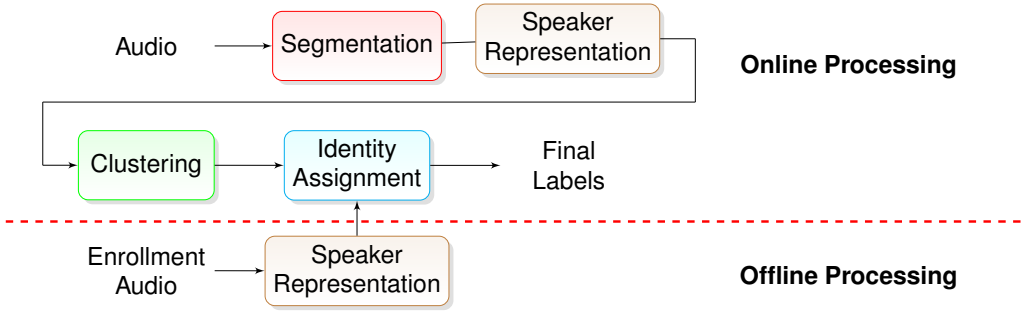

Figure 2. General block diagram for a speaker attribution system.

This diagram represents a Bottom-Up approach for speaker attribution. First, the given audio is divided into segments in which a single speaker is assumed. These segments are then transformed into representations that highlight the speaker discriminative properties. Taking into account these representations, the segments are clustered, grouping those with a common speaker. Finally, each one of these clusters are assigned to an identity, either an enrolled speaker or the generic unknown one. For this purpose, a side pipeline processes those audios belonging to the enrollment identities in an offline fashion. The described diagram has an alternative interpretation: A Bottom-Up diarization system isolates the speakers in the audio under analysis, assigning them to the corresponding identity afterwards. Regardless of the interpretation, most of these blocks are based on popular techniques from both speaker recognition and diarization.

The contribution of speaker recognition consists of tools to accurately represent the speakers by means of the embedding-backend paradigm. Thus, the speech of a speaker in a given piece of audio is transformed into a compact representation that highlights its discriminative properties. Some generative statistical alternatives are Joint Factor Analysis [1] or the well-known i-vectors [2]. With the advent of Deep Learning (DL), discriminative embeddings based on neural networks have also been proposed, such as x-vectors [4] and d-vectors [16]. On top, a backend is in charge of scoring how likely the speaker in the test audio is an enrolled speaker. As backend, the Probabilistic Linear

Discriminant Analysis (PLDA) [3] has been traditionally considered, although certain modifications, such as discriminative PLDA [17], the heavy-tailed PLDA [18,19] or the Neural PLDA [20], are now popular too. Unfortunately, all these techniques work under the assumption of a single speaker within the given recording; otherwise, the performance is severely degraded.

Due to the fact that speaker recognition techniques cannot deal with multiple speakers, diarization is in charge of isolating the audio from each speaker. In order to do so, diarization contributes with two stages: Segmentation identifies contiguous segments with a single speaker, and Clustering groups those segments that share a common speaker. Regarding the first block, the division of the audio is done by means of the Speaker Change Point Detection (SCPD) techniques. Some considered technologies for this subtask are metric-based (Bayesian Information Criterion (BIC) [21], ΔBIC [22], etc.) or models such as Hidden Markov Models (HMMs) [23] and Deep Neural Networks (DNNs) [24]. With respect to the clustering stage, multiple alternatives have been proposed, such as Agglomerative Hierarchical Clustering (AHC) [25,26], mean-shift [27–29], K-means [30,31] and Variational Bayes [32], or statistical approaches, such as i-vectors [33] or PLDA [34–38].

Finally, the identity assignment block is exclusive from speaker attribution and responsible for assigning each estimated cluster to the corresponding enrolled speaker. In order to make this assignment, clustering techniques are usually applied, such as the Ward method based on the Hotteling t-square statistic [39], the symmetrized Kullback–Leibler divergence [5,8] or the Cross-Likelihood Ratio (CLR) [7]. Other alternatives, i.e., the ones presented in [10], consider the assignment as a speaker recognition task. Alternatively, other contributions to speaker attribution present a totally new approach, integrating clustering and identity assignment blocks and simultaneously performing diarization and speaker attribution. This approximation as a dual task is solved by graph-based techniques [11].

Unfortunately, despite the great amount of developed techniques, the speaker attribution problem is still far from being solved for all domains. Broadcast data can be considered a wide collection of shows and genres with particular characteristics and, therefore, huge variability. This variability is due to the different acoustic environments in the data but also to other reasons related to the speakers (different number of speakers, unbalanced amount of speech among speakers, etc.). Thus, a general system may suffer from degradation to each particular scenario, and the same large amount of shows and genres discourages having multiple particular systems. Besides, in the broadcast domain, it is usually required to deal with unseen subdomains (e.g., a new show). While certain works have mitigated the domain mismatch between training and evaluation scenarios [37], in speaker attribution, we must also bear in mind the potential extra domains in the enrollment audios. This case is very common in real life, especially concerning public figures whose data can be collected in multiple conditions.

Finally, other important factors for the speaker attribution performance are the benefits and limitations of open-set and closed-set scenarios. The former scenario presents the unconstrained version of the speaker attribution task, whereas the latter condition simplifies the attribution problem by not considering the unknown identity, potentially boosting the performance. This consideration is important in the broadcast domain where some shows potentially fit into the closed-set condition, such as debate programs.

3. Experimental Protocol

3.1. Albayzín Corpus and Allowed Data

The dataset used for the evaluation of the systems is the Albayzín 2020 corpus (Evaluation plan available in http://catedrartve.unizar.es/reto2020/EvalPlan-SD-2020-v1.pdf, accessed on 13 September 2021). This dataset consists of approximately 750 h of broadcast content and is the accumulation of past Albayzin evaluations (2012, 2016, 2018 and 2020) proposed by Red Temática de Tecnologías del Habla (RTTH, http://rthabla.es, accessed on 13 September 2021). The content of the dataset, focused on the broadcast domain, considers multiple sources (3/24 TV channel, Corporación Aragonesa de Radio y Televisión (CARTV)

and Radio Televisión Española (RTVE)), media (radio and TV) as well as languages (Spanish and Catalan). The corpus integrates a wide range of shows and genres, from talk-shows to fiction. Albayzín evaluations are not exclusively focused on speaker technologies. Thus, a wide variety of metadata is released: around 180 h of human-annotated content, including time-referenced speaker labels, 100 h of audio labeled with coarse acoustic segmentation and 570 h of audio with the released broadcast subtitles.

For diarization and speaker attribution purposes, we only bear in mind the subset of audio with speaker labels, i.e., approximately 180 h of the total. In this analysis, we divide the corpus imitating the 2020 evaluation conditions: The audio from previous evaluations (2012, 2016 and 2018) is available for training purposes, consisting of approximately 140 h of content. The remaining audio, the data released in 2020, is divided into two subsets: development (4 h of audio) and test (37 h of content). The development set consists of two different shows: the documentary "Aquí la Tierra" as well as the teen drama "Bajo la Red". The test subset presents a larger heterogeneity with ten different shows that belong to four genres: Fiction series, such as the already mentioned "Bajo la Red", as well as "Boca Norte", "Si Fueras Tu" and "Wake-up"; two debate shows: "Los Desayunos de TVE" and "Millenium"; two comedy shows, namely "Ese Programa del que Usted me Habla" and "Neverfilms", are also evaluated. Finally, two magazines, "Aquí la Tierra" and "Comando Actualidad", complete the subset. (All involved shows are reproducible in RTVE Video on Demand (VoD) platform https://www.rtve.es/play. accessed on 13 September 2021) The selection of shows for the evaluation involves large variability in terms of duration (from 3.5 min up to 96 min) or the number of speakers (from 6 to 74). Furthermore, although certain shows ("Aquí la Tierra", "Bajo la Red" and "Millenium") are not exclusive to the test subset, this information was deliberately hidden by the organizers during the evaluation to avoid particular setups in these specific audios.

In order to perform the speaker attribution task, exclusive of the 2020 edition, extra audio for the enrolled speakers was also released. The provided data includes approximately an hour of audio involving 179 people (18 for development, 161 from test) in 193 recordings (25 for development and 168 in test). Hence, on average, each speaker is enrolled considering 20 s of audio, usually in a single condition.

The released data for Albayzín 2020 are originally configured according to an open-set scenario, with non-enrolled speakers in the evaluation audios as well as unseen enrolled speakers. A closed-set version can be obtained by masking all the speech from non-enrolled speakers in both online processing and evaluation.

In spite of the large amount of released data in the official Albayzín corpus, 2020 edition rules do not restrict the use of extra data if necessary. Hence, in our experiments, the Albayzín training subset is complemented with data from Video on Demand (VoD) by means of VoxCeleb 1 [40] and 2 [41]. We also make use of extra broadcast data with the manually annotated subsets from the Multi-Genre Broadcast (MGB) Challenge 2015 [42].

3.2. Performance Metrics for Diarization and Speaker Attribution

The original evaluation considers the Assignment Error Rate (AER) as the measure to evaluate the performance of the systems. This metric, conceived as a modification of the popular Diarization Error Rate (*DER*), reflects the proportion of misclassified audio.

The original *DER* metric assumes errors to be caused by three different causes: missed audio (lost audio with speech), false alarm audio (audio falsely considered to contain speech) and speaker error (speech assigned to a wrong speaker). Thus, the formulation of the *DER* metric is:

$$\begin{aligned} DER &= \frac{L_{\text{WRONG}}}{L} = \frac{L_{\text{MISS}} + L_{\text{FA}} + L_{\text{SPK}}}{L} \\ &= E_{\text{MISS}} + E_{\text{FA}} + E_{\text{SPK}} \end{aligned} \tag{1}$$

where L, L_{WRONG}, L_{MISS}, L_{FA} and L_{SPK} stand for the total length of the audio, the length of audio wrongly assigned and the lengths of audio per error cause, respectively, miss, false

alarm and speaker. Moreover, as shown in the formulation, *DER* can also be interpreted as a composition of three terms, E_{MISS}, E_{FA} and E_{SPK}, each one representing the specific error due to one of the previously described causes.

AER, heavily inspired by *DER*, follows a similar formulation that measures the proportion of misclassified audio. Furthermore, AER also differentiates among three causes of error: Miss, False Alarm and Speaker errors. However, differences with *DER* arise when defining these causes. Missed audio now reflects the lost audio exclusively coming from the enrollment speakers. The false alarm audio includes the audio falsely considered to contain an enrolled speaker. Finally, the speaker error is the audio attributed to an incorrect enrolled speaker.

4. Methodology

In this section, we describe the different alternatives under evaluation, all of which are based on the general speaker attribution block diagram illustrated in Figure 2 and rely on the embedding-PLDA paradigm. For this reason, we first describe all the common blocks along the systems, i.e., the front-end, the Speaker Change Point Detection (SCPD) and the embedding extraction. Then, we explain in detail the clustering block as well as the backend. Afterwards, we describe the identity assignment block. Next, we describe our first alternative, the direct assignment, which lacks diarization. The following alternative is the indirect assignment system, a similar system now performing diarization clustering. A new hybrid system simultaneously performing diarization and speaker attribution is studied afterwards. Our last alternatives under analysis are the semisupervised versions of the indirect assignment and hybrid systems. Finally, we also talk about the impact of open-set conditions on the systems under evaluation and how closed-set versions are built.

4.1. Front-End, SCPD and Embedding Extractor Blocks

The first block for all the systems under analysis is an MFCC [43] front-end. For a given audio Ω, a stream of 32 coefficient feature vectors is estimated according to a 25 ms window with a 15 ms overlap. No derivatives are considered.

Simultaneously, DNN-based Voice Activity Detection (VAD) labels are estimated [44]. The network, consisting of a joint work of convolutional layers and Long-Short Term Memory (LSTMs) [45], estimates a VAD label each 10 ms.

Both feature vectors and VAD labels are fed into the Speaker Change Point Detection (SCPD) block. This stage, dedicated to infer the speaker turn boundaries, makes use of ΔBIC [22], the differential form of BIC. This estimation works in terms of a 6-second sliding window, in which we assume there is, at most, a speaker turn boundary. Each involved speaker in the analysis is modeled by means of a full-covariance Gaussian distribution. Besides, the VAD labels delimit the parts of the audio in which the analysis is performed. In the given data, the described configuration provides segments of an approximately 3-second length on average.

The identified segments are converted into embeddings using a modification of the extended x-vector [46] based on Time Delay Neural Networks (TDNNs) [47]. The modification is the inclusion of multi-head self-attention [48] in the pooling layer. This network, trained on VoxCeleb1 and 2, provides embeddings of dimension 512. These embeddings undergo centering and LDA whitening (reducing dimension to 200), both trained with MGB as well as the Albayzín training subset, and finally length normalization [49]. These embeddings will be referred to as $\mathbf{\Phi}$. A similar extraction pipeline working offline is in charge of the enrollment audios. The enrollment embeddings will be named $\mathbf{\Phi}_{enroll}$.

4.2. PLDA Tree-Based Clustering Block

The considered clustering block relies on the idea originally proposed in [36]. It uses a model of the PLDA family to estimate the set of N labels $\Theta = [\theta_1, \ldots, \theta_N]$ that best explain the set of N embeddings $\mathbf{\Phi} = [\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_N]$. Thus, the label θ_j indicates the cluster to

which the embedding $\boldsymbol{\phi}_j$ belongs. The proposed approach works based on the Maximum a Posteriori (MAP) estimation of the labels:

$$\Theta_{clust} = \arg\max_{\Theta} P(\Theta|\boldsymbol{\Phi}) = \arg\max_{\Theta} \frac{P(\boldsymbol{\Phi},\Theta)}{P(\boldsymbol{\Phi})} \tag{2}$$

Working in terms of the latter expression, the only term depending on the labels is $P(\boldsymbol{\Phi},\Theta)$, which can be decomposed according to the product rule of probability as:

$$P(\boldsymbol{\Phi},\Theta) = \prod_{j=2}^{N} P\left(\boldsymbol{\phi}_j,\theta_j|\boldsymbol{\phi}_1^{j-1},\theta_1^{j-1}\right) P(\boldsymbol{\phi}_1,\theta_1) \tag{3}$$

respectively, being $\boldsymbol{\phi}_1^{j-1}$ and θ_1^{j-1} subsets from $\boldsymbol{\Phi}$ and Θ with the first $j-1$ elements $\left(\boldsymbol{\phi}_1^{j-1} = \left[\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_{j-1}\right]\right.$ and $\left.\theta_1^{j-1} = [\theta_1,\ldots,\theta_{j-1}]\right)$.

Inspired by [34], we can make the term $P\left(\boldsymbol{\phi}_j,\theta_j|\boldsymbol{\phi}_1^{j-1},\theta_1^{j-1}\right)$ more tractable by decomposing it into a conditional distribution depending on the embeddings and a prior distribution for the labels:

$$P\left(\boldsymbol{\phi}_j,\theta_j|\boldsymbol{\phi}_1^{j-1},\theta_1^{j-1}\right) = P\left(\boldsymbol{\phi}_j|\theta_j,\boldsymbol{\phi}_1^{j-1},\theta_1^{j-1}\right) P\left(\theta_j|\theta_1^{j-1}\right) \tag{4}$$

This decomposition isolates a term $P\left(\boldsymbol{\phi}_j|\theta_j,\boldsymbol{\phi}_1^{j-1},\theta_1^{j-1}\right)$ in which the embedding j depends on its jth label as well as previous embeddings and labels. As the design choice, we define it as a mixture of simpler distributions controlled by the means of the variable θ_j. We define this variable as a one-hot sample with I values ($\theta_j = \{\theta_{1j},\ldots,\theta_{ij},..\theta_{Ij}\}$), where I is the number of candidate clusters. Hence, the ith component of θ_j takes the value of one exclusively if the embedding j belongs to cluster i. Additionally, we also impose the restriction that the embedding j, when belonging to the cluster i, should be exclusively explained by those embeddings already assigned to this cluster. This subset of previous embeddings assigned to cluster i at time j is denoted as $\boldsymbol{\Phi}_{ij}$. Under these conditions:

$$P\left(\boldsymbol{\phi}_j|\theta_j,\boldsymbol{\phi}_1^{j-1},\theta_1^{j-1}\right) = \prod_{i=1}^{I} P\left(\boldsymbol{\phi}_j|\boldsymbol{\Phi}_{ij}\right)^{\theta_{ij}} \tag{5}$$

As we described before, we want this model to be based on the PLDA family. This family of models makes use of a latent variable $\mathbf{y}_i$ that is shared by all elements from the same cluster. Consequently, we can redefine $P\left(\boldsymbol{\phi}_j|\boldsymbol{\Phi}_{ij}\right)$ as:

$$P\left(\boldsymbol{\phi}_j|\boldsymbol{\Phi}_{ij}\right) = \int P\left(\boldsymbol{\phi}_j|\mathbf{y}_i\right) P(\mathbf{y}_i|\boldsymbol{\Phi}_{ij}) d\mathbf{y}_i \tag{6}$$

By going backwards, we have obtained two familiar distributions when dealing with the PLDA family: $P\left(\boldsymbol{\phi}_j|\mathbf{y}_i\right)$ and $P(\mathbf{y}_i|\boldsymbol{\Phi}_{ij})$. In our formulations, we consider our model to be a Fully Bayesian PLDA (FBPLDA) [50] with a single latent variable. This choice implies:

$$P\left(\boldsymbol{\phi}_j|\mathbf{y}_i\right) \sim \mathcal{N}\left(\boldsymbol{\phi}_j|\boldsymbol{\mu} + \mathbf{V}\mathbf{y}_i, \mathbf{W}^{-1}\right) \tag{7}$$

where $\boldsymbol{\mu}$ is the speaker independent term, $\mathbf{V}$ a low rank matrix describing the speaker subspace and $\mathbf{W}$ a full rank matrix explaining the intra-speaker variability space. Similarly, the definition of $P(\mathbf{y}_i|\boldsymbol{\Phi}_{ij})$ according to the same model is:

$$P(\mathbf{y}_i|\boldsymbol{\Phi}_{ij}) \sim \mathcal{N}\left(\boldsymbol{\phi}_j|\boldsymbol{\mu}_{\mathbf{y}_i}(j), \mathbf{L}_{\mathbf{y}_i}(j)\right) \tag{8}$$

$$\mathbf{L}_{\mathbf{y}_i}(j) = \mathbf{I} + \mathbf{V}^T \sum_{k=1}^{j-1} \theta_{ik} \mathbf{W}\mathbf{V} \tag{9}$$

$$\boldsymbol{\mu}_{\mathbf{y}_i}(j) = \mathbf{L}_{\mathbf{y}_i}(j)^{-1} \mathbf{V}^T \mathbf{W} \sum_{k=1}^{j-1} \theta_{ik}(\boldsymbol{\phi}_k - \boldsymbol{\mu}) \tag{10}$$

In addition, for the well-known closed form definitions, the choice for the FBPLDA model also makes the previously mentioned integral have a closed-form solution:

$$P\left(\boldsymbol{\phi}_j|\boldsymbol{\Phi}_{ij}\right) \sim \mathcal{N}\left(\boldsymbol{\phi}_j|\boldsymbol{\mu}_i(j), \boldsymbol{\Sigma}_i(j)\right) \tag{11}$$

$$\boldsymbol{\mu}_i(j) = \boldsymbol{\mu} + \mathbf{V}\boldsymbol{\mu}_{\mathbf{y}_i}(j) \tag{12}$$

$$\boldsymbol{\Sigma}_i(j) = \mathbf{W}^{-1} + \mathbf{V}\mathbf{L}_{\mathbf{y}_i}^{-1}(j)\mathbf{V}^T \tag{13}$$

With respect to the label prior $P\left(\theta_j|\theta_1^{j-1}\right)$, we opt for its definition applying the same distribution as in [51]. In this work, the authors make use of a modification of the Distance Dependent Chinese Restaurant (DDCR) process [52]. This process explains the occupation of an infinite series of clusters by a sequence of elements, assigning new elements to already existing clusters or creating a new group.

The presented decomposition fits a unique tree structure in which nodes stand for assignment decisions, and leaves, all at depth N, represent the possible partitions Θ for the set of embeddings $\boldsymbol{\Phi}$. Due to the high complexity of finding the optimal set of labels Θ [53], this tree clustering approach follows a suboptimal iterative inference of the labels: We estimate the best partition θ_1^j for the subset $\boldsymbol{\phi}_1^j$ given a solution for a simplified problem and the set of labels θ_1^{j-1} for the subset of embeddings $\boldsymbol{\phi}_1^{j-1}$. This procedure is complemented by the M-algorithm [54], which simultaneously tracks the best M alternative partitions at each time.

In our experiments, we opt to make use of a Fully Bayesian PLDA of dimension 100 and trained with all the available broadcast data, i.e., MGB and previous editions of Albayzín evaluations.

4.3. The Identity Assignment Block

The considered identity assignment block consists of a speaker recognition architecture, composed of the same clustering PLDA model, followed by score normalization and calibration stages. First, the previously mentioned PLDA backend, by means of its likelihood ratio, scores how likely a subset of embeddings j from the audio under analysis with a common speaker resemble the enrollment speaker i. These scores s_{ij} are then normalized by means of the adaptive S-norm [55]. The normalization cohort consists of MGB 2015 labeled data. Finally, the normalized scores are calibrated according to a threshold γ prior to the decision making. Those scores overcoming the threshold are considered the *target*, i.e., the test speaker and the enrollment speaker are the same person. This threshold is adjusted by AER minimization in terms of a calibration subset $\boldsymbol{\Phi}_{\text{calib}}$ and the enrollment embeddings $\boldsymbol{\Phi}_{\text{enroll}}$ as follows:

$$\gamma = \arg\min_{\gamma}(\text{AER}(\boldsymbol{\Phi}_{\text{calib}}, \boldsymbol{\Phi}_{\text{enroll}}, \gamma)) \tag{14}$$

where $\boldsymbol{\Phi}_{\text{calib}}$ and $\boldsymbol{\Phi}_{\text{enroll}}$ represent the set of embeddings from calibration as well as the enrollment speakers. The last step in the identity assignment block is the decision making.

Any subset of embeddings is assigned to the enrolled identity with the highest score only if this value overcomes the calibration threshold, being assigned to the generic unknown identity otherwise. Mathematically, the assigned identity (θ_j) for a subset of embeddings j with respect to the enrolled identity i is:

$$\theta_j = \begin{cases} \arg\max_i(s_{ij}|s_{ij} > \gamma) & \text{if } \exists i|s_{ij} > \gamma \\ \text{Unknown} & \text{if } \forall i, s_{ij} < \gamma \end{cases} \quad (15)$$

where s_{ij} stands for the normalized PLDA log-likelihood ratio score between the embedding j and the enrolled identity i.

4.4. The Direct Assignment Approach

The first proposal follows a traditional out-of-the box speaker recognition architecture. This architecture fits the diagram block from Figure 2 except for the clustering stage, which is not applied. Additionally, the identity assignment block follows the previously described speaker recognition philosophy. In order to explain it in detail we present Figure 3, which illustrates its functionality. Given the analysis audio, we individually assign an identity, enrolled or generic unknown, to each embedding $\boldsymbol{\phi}_i \in \boldsymbol{\Phi}$ representing an audio segment. During the assignment process, each score s_{ij} stands for the similarity between the individual embedding j from $\boldsymbol{\Phi}$ compared to the enrollment speaker i. Moreover, the role of the calibration subset $\boldsymbol{\Phi}_{\text{calib}}$ is played by those embeddings from the development subset. This same calibration is also applied to the test subset. The detailed flowchart is illustrated in Figure 4. Given the audio under analysis Ω, we perform its feature extraction, segment the estimated information (VAD and SCPD stages are involved) and extract the set of embeddings per segment (red box). These embeddings are compared to those obtained from the enrollment audios Ω_{enroll} (yellow box), processing the features for the whole recording. Both sets of embeddings are fed into the identity assignment block to obtain the final labels.

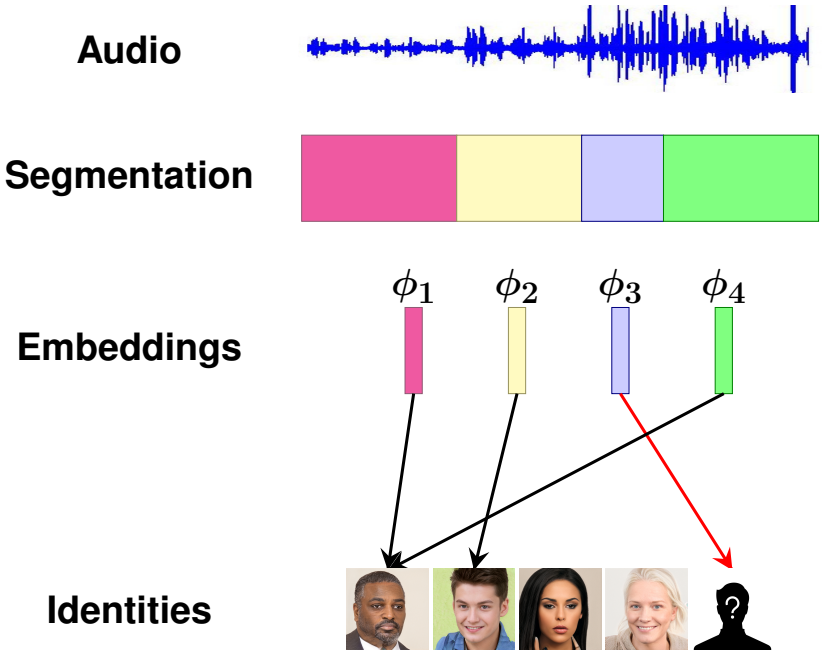

Figure 3. Diagram of the direct assignment approach. The embeddings obtained from the different parts of the given audio are independently assigned to the identities. These assignments can be done to enrolled identities or the generic unknown one (red arrow).

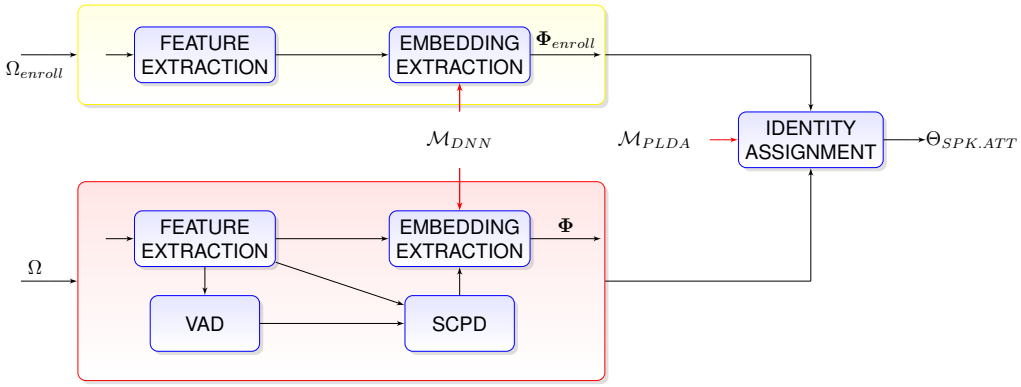

Figure 4. Flowchart of the direct assignment approach. Red and yellow boxes, respectively, represent the embedding extraction pipelines for the evaluation audio Ω (online) and enrollment audios Ω_{enroll} (offline). The obtained embeddings ($\mathbf{\Phi}$ and $\mathbf{\Phi}_{enroll}$) are taken into account in the identity assignment block.

4.5. Clustering and Assignment: The Indirect Assignment Approximation

One of the great limitations of the previous alternative is the robustness of embeddings when obtained from a low amount of audio [56]. Thus, decisions based on more audio should be more reliable.

Thus, the next approach, the indirect assignment, straightforwardly follows the diagram block described in Figure 2, also applying the speaker recognition identity assignment block. The functionality of the system is described in Figure 5. Similarly to the direct assignment block, we obtain the embeddings $\mathbf{\Phi}$ according to the segments estimated from the given audio. However, compared to the direct assignment alternative, this system performs clustering prior to the assignment stage. Afterwards, the identity assignment subtask is performed similarly to the direct assignment approach but with an important difference: While in its predecessor system, the score s_{ij} compared the embedding j with the enrolled identity i, but in this new version, s_{ij} compares all embeddings assigned to the cluster $\mathcal{C}_j$ with the ith speaker of interest. Again, we consider the calibration subset $\mathbf{\Phi}_{\text{calib}}$ role to be played by the development subset. This learned calibration is then used with the test subset. Figure 6 shows the flowchart for this approximation.

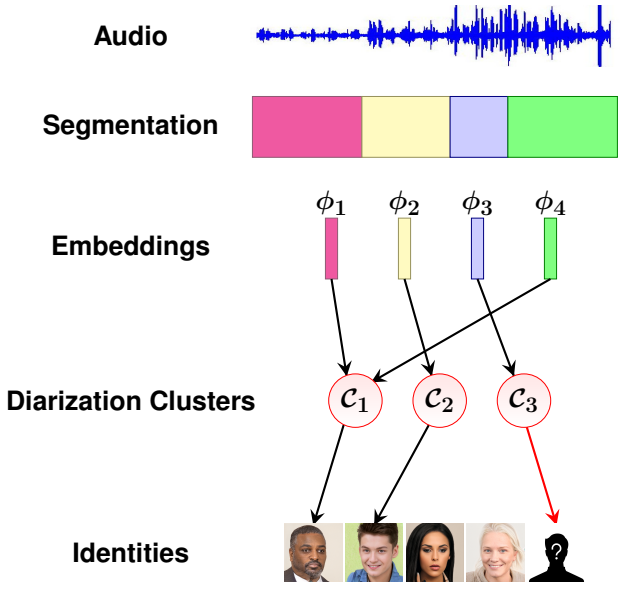

Figure 5. Diagram of the indirect assignment approach. Embeddings from the audio are first clustered during diarization ($\mathcal{C}_1$,...,$\mathcal{C}_3$). Then, clusters are assigned to the available identities, either the enrolled speakers or the unknown generic cluster (red arrow).

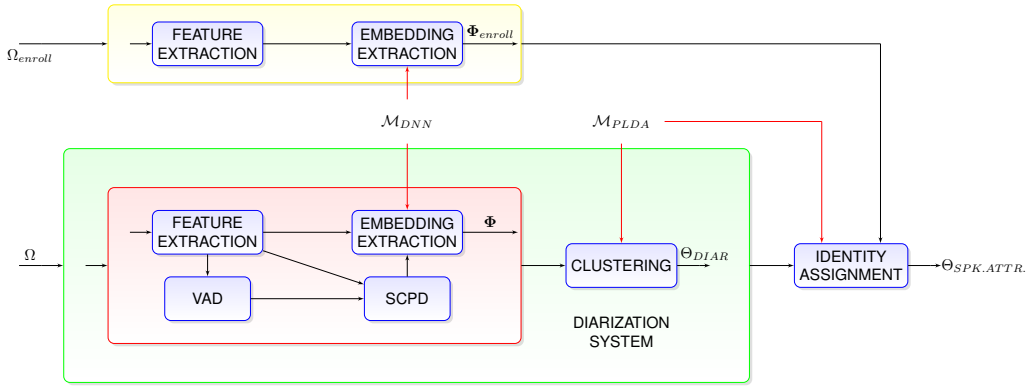

Figure 6. Flowchart of the indirect assignment approach. Red and yellow boxes, respectively, stand for embedding extraction pipelines for the evaluation audio Ω (online) and enrollment audios Ω_{enroll}. The green box means a diarization system, which clusters the evaluation embeddings $\mathbf{\Phi}$ to obtain diarization labels Θ_{DIAR}. The obtained embeddings ($\mathbf{\Phi}$ and $\mathbf{\Phi}_{enroll}$) as well as the estimated labels Θ_{DIAR} are taken into account in the identity assignment block.

Our system starts with the embedding extraction from the audio under analysis Ω (red box). This process involves a feature extraction of the audio followed by its segmentation (VAD and SCPD stages) and finally the embedding estimation. The estimated embeddings are fed into the clustering block to obtain the diarization labels. Meanwhile, enrollment audios Ω_{enroll} undergo a similar embedding extraction pipeline. Then, the identity assignment block is responsible for estimating the final labels according to all the estimated information, i.e., the two sets of embeddings as well as the diarization labels.

The choice for clustering addition implies a tradeoff: While clustering may boost the system performance, it increases the computational cost and also adds a potential performance degradation due to cluster impurities. Regarding other decisions, as a design choice, we do not exclude the assignment of multiple diarization clusters to a single enrolled speaker. This decision is made under the assumption that we can fix diarization errors.

4.6. Hybrid Solution

The next alternative, inspired by [11], presents a new approach that simultaneously performs diarization and speaker attribution using the enrollment audio as anchors. For this purpose, we consider a modification of the algorithm proposed in [36] by merging the clustering and identity assignment blocks. The system, whose functionality is represented in Figure 7, starts estimating the embeddings obtained from the inferred segments. Regarding the new algorithm, it is a statistical Maximum A Posteriori (MAP) solution to estimate the set of labels $\Theta' = \Theta_{enroll} \cup \Theta$ that best explains the set of embeddings $\mathbf{\Phi}' = \mathbf{\Phi}_{enroll} \cup \mathbf{\Phi}$, i.e., the union of enrollment and evaluation embeddings, assuming the labels Θ_{enroll} for the enrollment subset are fixed. This enrollment information generates a set of anchor clusters for the speakers of interest. Thus, the algorithm iteratively assigns the embeddings from the audio under evaluation to the anchor clusters or new clusters created on the fly for speakers out of the group of people of interest. The flowchart for the presented solution is shown in Figure 8. From the evaluation audio Ω, a stream of features is extracted and used for segmentation (VAD and SCPD blocks), extracting one embedding per estimated segment (red box). Similarly, enrollment audios are processed to obtain their embeddings in the yellow box. Finally, both sets of embeddings are fed into the new hybrid clustering and identity assignment block (green box) to obtain the final labels.

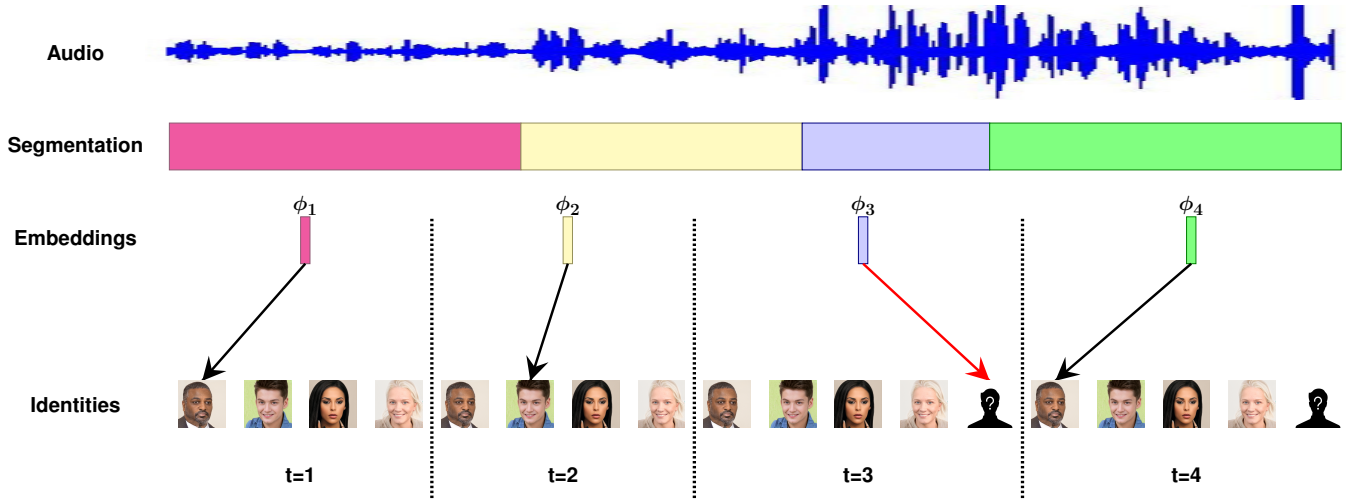

Figure 7. Diagram of the hybrid approach. Embeddings ($\boldsymbol{\phi}_1$,...,$\boldsymbol{\phi}_4$) are sequentially assigned to the available clusters at each time t. Initially, the available clusters are only those for the enrolled speakers. When the embedding is not assigned to an existing cluster (t=3), it is responsible for an extra cluster for an unknown speaker (red arrow). This new cluster is then available along the posterior assignments.

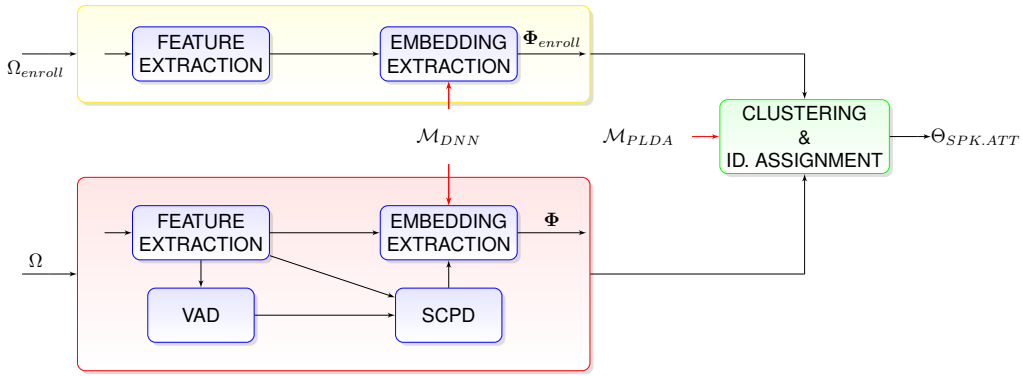

Figure 8. Flowchart for the hybrid approach. Red and yellow boxes, respectively, stand for the embedding extraction pipelines for the evaluation audio Ω (online) and the enrollment audios Ω_{enroll} (offline). Both sets of embeddings are used in the new hybrid clustering and identity assignment block.

4.7. Semisupervised Alternative

In the semisupervised alternative, we consider an scenario where rather than blindly dealing with any domain, we can previously obtain a small portion of data with manual annotation. Thanks to these data, we optimize the system to the domain under analysis. The main drawback for this option is the cost of data gathering and annotation, creating a trade-off between potential performance benefits and cost. In this line of research, we propose two different semisupervised options developed around previously described systems from Sections 4.5 and 4.6. These systems take into account three types of input embeddings: $\boldsymbol{\Phi}$,$\boldsymbol{\Phi}_{\text{enroll}}$ and $\boldsymbol{\Phi}_{\text{in}}$, respectively, obtained from the audio under analysis, the enrollment audios and the new labeled in-domain data.

The first semisupervised option is a modification of the system described in Section 4.5 and illustrated in Figure 9. In this system, we modify the subset of embeddings playing the role of $\boldsymbol{\Phi}_{\text{calib}}$ for the threshold adjustment. While in the unsupervised system, this role is played by the whole development subset with potentially out-of-domain data, we now make use of the annotated in-domain data $\boldsymbol{\Phi}_{\text{in}}$.

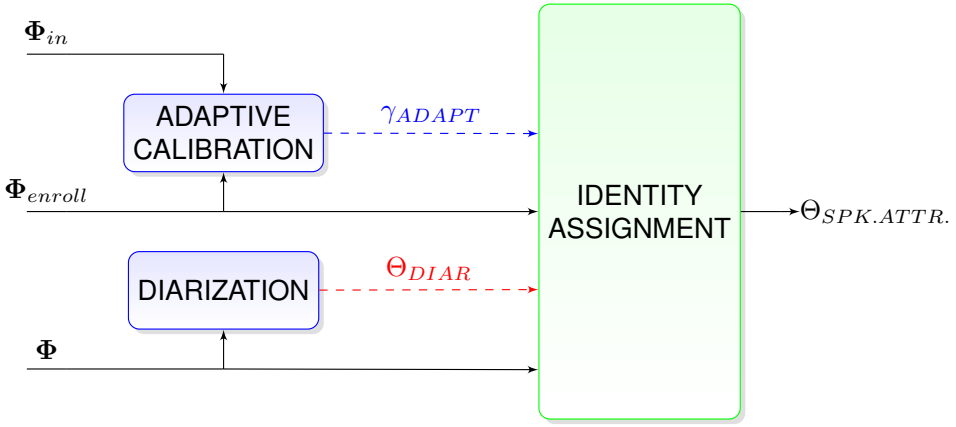

Figure 9. Diagram block for the semisupervised approach based on the indirect assignment approach.

The second semisupervised alternative works with the system from Section 4.6. This alternative modifies the definition of anchor embeddings. While the unsupervised hybrid system builds the anchor clusters exclusively based on the enrollment embeddings $\mathbf{\Phi}_{\text{enroll}}$, the semisupervised version now considers the joint work of enrollment embeddings as well as those from the new labeled in-domain data $\mathbf{\Phi}_{enroll} \cup \mathbf{\Phi}_{in}$. By doing so, anchor clusters are reassured to contain traces of in-domain audio regardless of the enrollment audio nature.

In both presented systems, the influence of the new in-domain data is restricted to the assignment algorithms, keeping the backend models unmodified. This choice is made despite the potential benefits of model adaptation to make a fair comparison between unsupervised and semisupervised styles.

4.8. Open-Set vs. Closed-Set Conditions

As we previously described, depending on the data under analysis, we can consider two different modalities: closed-set and open-set. The only difference is that the former modality assumes that all speakers in the audio under analysis are enrolled in the system, while the latter one does not, providing a more general solution that is more complex to deal with.

While all the previously described speaker attribution systems work under the assumption of open-set conditions, we can include small modifications to exploit the closed-set simplifications. The direct and indirect assignment systems as well as the semisupervised modification can be transformed by not including any calibration block and thus making the assignment to the enrollment speaker with highest score. This simplification obeys the fact that the calibration threshold is used to assign audio to the generic unknown speaker, which is non-existent in the closed-set scenario. By contrast, those systems based on the hybrid approach (either in unsupervised or semisupervised fashion) are insensitive to the closed-set condition, thus not requiring any modification.

5. Results

The goal of this work is to provide a detailed analysis of the speaker attribution problem in the complex broadcast domain. However, due to the high complexity of the task, we have divided it into the following three subtasks:

- An illustration of the influence of diarization on the speaker attribution problem;
- A depiction of the impact of broadcast domain variability into the speaker attribution task;
- A proposal of alternative approximations to deal with this variability, with special emphasis on unseen domains.

5.1. The Influence of Diarization

In our first set of experiments, we evaluate our two proposed traditional alternatives: the direct and the indirect assignment systems. While the former system only works in terms of isolated segments, the latter performs clustering to obtain diarization labels used during the identity assignment. The quality of these diarization labels is shown in Table 1, where both development and test subsets from Albayzín 2020 are evaluated. The analysis is extended to both closed-set and open-set conditions.

Table 1. DER (%) results for the Albayzín 2020 corpus, including results for both development and test subsets.

Scenario	Development DER (%)	Test DER (%)
Closed scenario	6.72	8.67
Open scenario	17.27	15.16

The results in Table 1 show interesting performances in all the experiments but far from a perfect estimation of the labels. The diarization system shows a similar behavior between development and test regardless of the scenario, showing robustness against domain mismatch. We also see how the restrictions in the audio under analysis, i.e., the closed set assumption, implies important improvements in performance (up to a relative 61%).

Once the diarization system is characterized, we compare the two systems in the speaker attribution task. In order to exclusively analyze the impact of diarization, we get rid of the domain issues by applying adaptive oracle calibrations to the data. These calibrations are obtained by the previously described criteria, only substituting the subset playing the role $\mathbf{\Phi}_{\text{calib}}$, which is now played by the same data under analysis. This oracle calibration has been estimated in three different degrees of generality: a calibration for all the audios in the development and test subsets (subset-level), an individual treatment per show (show-level) and a particular calibration per audio (audio-level). The obtained results for this experiment are included in Table 2.

Table 2. Study of the impact of diarization on speaker attribution with oracle calibration. Experiments carried out with direct (without diarization) and indirect assignment (with diarization) systems. Three degrees of calibration generality are shown. AER (%) results for the Albayzin 2020 development and test subsets. Experiment corresponding to the open condition.

Data Subset	Subset-Level		Show-Level		Audio-Level	
	Direct	Indirect	Direct	Indirect	Direct	Indirect
Dev. subset	41.91	**37.45**	41.27	**35.88**	39.89	**29.09**
Eval. subset	48.19	**34.87**	41.70	**28.10**	40.31	**26.54**

The results illustrated in Table 2 indicate significant benefits (from a relative 27% up to a 34% improvement) when diarization is applied rather than considering individual segments. Thus, diarization clusters, despite their impurities, provide robustness to those segments with a low amount of audio. Furthermore, we also see the importance of specificity in calibration. The more particular the calibration, the larger potential improvements can be achieved, although resources must be more specific for the particular domain.

5.2. Broadcast Domain Mismatch in Speaker Attribution

The results from Table 2 have shown the importance of diarization as well as the impact of individual adjustments of the system to each specific piece of information (subset, show or audio). However, when no oracle calibration is available, systems are likely to degrade.

The next experiment analyses the two previously evaluated systems, i.e., the direct assignment as well as the indirect assignment one. However, now, we do not apply an oracle

calibration but estimate one according to the whole development subset. Additionally, we also evaluate the hybrid system from Section 4.6, which is supposedly more robust against domain mismatches. The obtained results for these three systems in both closed-set and open-set conditions are included in Table 3.

Table 3. AER (%) results for the Albayzín 2020 corpus. Results of direct and indirect assignment as well as the hybrid systems, including results for both development and test subsets. Experiment corresponding to closed and open conditions.

Subset	Closed Condition			Open Condition		
	Direct	Indirect	Hybrid	Direct	Indirect	Hybrid
Dev. subset	**13.73**	15.27	15.89	41.91	**37.45**	37.68
Eval. subset	25.11	17.20	**16.49**	65.31	60.34	**31.95**

The results in Table 3 reveal that systems depending on a tuned calibration (direct and indirect assignment) strongly suffer from degradation with respect to the oracle result in Table 2 (above a relative higher than a 35% relative degradation) when dealing with domain mismatch. Moreover, we also see great differences in behavior between development and test (degradations over a relative 55%), while particular calibrations (Table 2) did not. Furthermore, this loss of performance is characteristic for the open-set condition; loss of performance does appear in the closed set where no calibration is applied. Regarding the hybrid system, its behavior manages to obtain the best results in the open-set condition, obtaining improvements up to a relative 47%, which also outperforms some scenarios with oracle calibration in Table 2. When applied to the closed-set scenario, its performance is also the best one, yet the improvements are not as noticeable (relative 4%) compared to its direct and indirect assignment counterparts. This is because a closed-set scenario does not require any calibration tuning to fit the audio under analysis.

5.3. Semisupervised Solutions

In the previous experiments, we have confirmed the alternative definition of broadcast as a collection of audios with particular characteristics. These individual properties generate an important domain mismatch that can cause significant harmful effects on the speaker attribution task.

While, in the previous experiments, we developed an unsupervised technique, our hybrid system, robust enough to deal with unseen scenarios, we also want to evaluate other types of solutions. One of them are semisupervised alternatives, also known as human-assisted solutions, which require available small portions of annotated in-domain data.

Bearing in mind those results from Table 2, we consider all the audios from the same show as the domain. This choice is done according to the trade-off between specificity and simplicity to gather the data. Then, for each show under analysis, we consider the audio and speaker labels of an episode available as adaptation in-domain data.

Due to the way Albayzín 2020 data subsets are arranged, most shows are exclusive from a subset. Hence, we must alter the Albayzín 2020 subsets in order to reassure this annotated adaptation audio per show. The considered modification divides each subset, development and test, into two parts: The first part consists of the annotated audios, one per show in the subset, for adaptation purposes. In our experiments, we consider the first episode in chronological order for adaptation. The second part of the subset, with the remaining episodes, is considered as the new evaluation subset.

The next experiments evaluate the indirect assignment system as well as the hybrid one in both unsupervised and human-assisted manners with the new subsets. The results of these experiments are contained in Table 4.

Table 4. AER (%) results for the Albayzín 2020 corpus for the assisted configuration. Results from indirect assignment and hybrid systems, including results for both development and test subsets. Experiment corresponds to an open-set condition.

Data Subset	Unsupervised		Semisupervised	
	Indirect	Hybrid	Indirect	Hybrid
Dev. subset	**38.86**	39.07	42.40	**38.45**
Eval. subset	59.00	**30.56**	30.66	**28.74**

The results in Table 4 indicate the benefits of small available portions of annotated in-domain data. By only having a single hour or audio manually annotated, systems previously seen as weak against domain mismatch (such as the indirect assignment system) now obtain significant improvements (a relative 48%). Actually, the same indirect system, now with an adapted calibration, manages to obtain results similar to those obtained with oracle calibration for shows. With respect to the hybrid system, the obtained benefits are not as noticeable (a relative 6% improvement) but are the best results obtained in the whole study.

6. Conclusions

Through this work, our goal was the improvement of the speaker attribution problem when dealing with broadcast data. These data are characterized for their great variability—also defined as a collection of particular domains with individual characteristics. For speaker attribution improvement, we studied three subtasks: the impact of diarization in the results, the importance of domain mismatch in the approaches and the proposal of alternative approximations that are robust enough to manage unseen scenarios.

With respect to the importance of diarization, our experiments with two straightforward approaches (direct and indirect assignment systems) confirm the benefits of using diarization labels (approximately a relative 34% improvement) despite being noisy (17% DER on the same dataset) when domain issues are canceled. Thus, the accumulation of acoustic information thanks to diarization significantly compensates the poor individual robustness of each obtained embedding due to its short length (around 3 s).

However, domain issues are a real problem in broadcast data. The same experiments with adaptive oracle calibrations show improvements up to a relative 24% by considering more specific domains. By contrast, real calibration adjusted during development can significantly degrade the performance due to domain mismatch, in our case, up to a relative 73%.

To solve it, we have proposed two alternatives: A new hybrid system that simultaneously performs diarization and speaker attribution as well as a semisupervised approach making use of a limited amount of annotated in-domain data. Regarding our hybrid alternative, it requires no calibration, showing great robustness against domain mismatch and obtaining relative improvements of 47% compared to the traditional counterparts. Moreover, this new approach has managed to overcome the results obtained with some adaptive oracle calibrations.

With respect to the semisupervised approach, it manages to improve the performance in both unsupervised proposals, indirect and the new hybrid system. These improvements are particularly interesting in the indirect assignment system, which only needed a small portion of in-domain audio to boost its performance (relative 48% improvement). With respect to the hybrid system, its semisupervised version offers a much more reduced improvement (a relative 6%) but obtains the best results with the dataset.

Author Contributions: Conceptualization, I.V. and A.O.; methodology, I.V.; software, I.V.; validation, I.V.; formal analysis, I.V. and A.O.; investigation, I.V. and A.O.; resources, A.O., A.M. and E.L.; data curation, A.O. and E.L.; writing—original draft preparation, I.V.; writing—review and editing, I.V. and A.O.; visualization, I.V., A.O., A.M. and E.L.; supervision, A.O., A.M. and E.L.; project

administration, A.O., A.M. and E.L.; funding acquisition, A.O., A.M. and E.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie Grant 101007666; in part by FEDER/Spanish Ministry of Economy and Competitiveness under Grant TIN2017-85854-C4-1-R, and in part by the Government of Aragón under Grant Group T36_20R

Data Availability Statement: RTVE data are available upon request through http://catedrartve.unizar.es/albayzin.html accessed on 13 September 2021.

Acknowledgments: We gratefully acknowledge the support of the NVIDIA Corporation with the donation of a Titan V GPU. This material is based upon work supported by Google Cloud.

Conflicts of Interest: The authors declare no conflict of interest. The founders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AER	Assignment Error Rate
AHC	Agglomerative Hierarchical Clustering
ASR	Automatic Speech Recognition
CARTV	Corporación Aragonesa de Radio y Televisión
CLR	Cross Likelihood Ratio
DER	Diarization Error Rate
JFA	Joint Factor Analysis
LSTM	Long-Short Term Memory
MAP	Maximum a Posteriori
MGB	Multi-Genre Broadcast
PLDA	Probabilistic Linear Discriminant Analysis
RTTH	Red Temática de Tecnologías del Habla
RTVE	Radio Televisión Española
SCPD	Speaker Change Point Detection
TDNN	Time Delay Neural Network
VAD	Voice Activity Detection

References

1. Kenny, P. Joint Factor Analysis of Speaker and Session Variability: Theory and Algorithms. In *(Report) CRIM-06/08-13*; CRIM: Montreal, QC, Canada, 2005; pp. 1–17.
2. Dehak, N.; Kenny, P.; Dehak, R.; Dumouchel, P.; Ouellet, P. Front-End Factor Analysis For Speaker Verification. *IEEE TASLP* **2011**, *19*, 788–798, doi:10.1109/TASL.2010.2064307.
3. Prince, S.J.D.; Elder, J.H. Probabilistic Linear Discriminant Analysis for Inferences About Identity. In Proceedings of the IEEE International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–21 October 2007; pp. 1–8
4. Snyder, D.; Ghahremani, P.; Povey, D.; Garcia-Romero, D.; Carmiel, Y.; Khudanpur, S. Deep Neural Network-based Speaker Embeddings for End-to-end Speaker Verification. *IEEE SLT* **2016**, 165–170, doi:10.1109/SLT.2016.7846260.
5. Ferras, M.; Boudard, H. Speaker diarization and linking of large corpora. *IEEE SLT* **2012**, 280–285, doi:10.1109/SLT.2012.6424236.
6. Ghaemmaghami, H.; Dean, D.; Sridharan, S. Speaker Linking Using Complete-Linkage Clustering. In Proceedings of the 14th Australasian International Conference on Speech Science and Technology, Sydney, Australia, 3–6 December 2012; pp. 1–4
7. Ghaemmaghami, H.; Dean, D.; Sridharan, S.; van Leeuwen, D.A. A study of speaker clustering for speaker attribution in large telephone conversation datasets. *Comput. Speech Lang.* **2016**, *40*, 23–45, doi:10.1016/j.csl.2016.03.005.
8. Ferras, M.; Madikeri, S.; Bourlard, H. Speaker Diarization and Linking of Meeting Data. *IEEE ACM TASLP* **2016**, *24*, 1935–1945, doi:10.1109/TASLP.2016.2590139.
9. Ferras, M.; Madikeri, S.; Motlicek, P.; Bourlard, H. System Fusion and Speaker Linking for Longitudinal Diarization of TV Shows. *IEEE ICASSP* **2016**, 5495–5499. doi:10.1109/ICASSP.2016.7472728.
10. Viñals, I.; Gimeno, P.; Ortega, A.; Miguel, A.; Lleida, E. Diarization and Identity Assignment Compatibility in the Albayzín 2020 Challenge. *Iberspeech* **2021**, 94–98. doi:10.21437/IberSPEECH.2021-20.
11. Wang, J.; Xiao, X.; Wu, J.; Ramamurthy, R.; Rudzicz, F.; Brudno, M. Speaker attribution with voice profiles by graph-based semi-supervised learning. *arXiv* **2020**, arXiv:2102.03634.

12. Lleida, E.; Ortega, A.; Miguel, A.; Bazán, V.; Pérez, C.; Gómez, M.; de Prada, A. Albayzin 2018 evaluation: The IberSpeech-RTVE challenge on speech technologies for Spanish broadcast media. *Appl. Sci.* **2019**, *9*, 1–22, doi:10.3390/app9245412.
13. van Leeuwen, D. Speaker Linking in Large Data Sets. In Proceedings of the Speaker and Language Recognition Workshop, ODYSSEY, Brno, Czech Republic, 28 June–1 July 2010; pp. 202–208.
14. Hujibregts, M.; van Leeuwen, D.A. Large-Scale Speaker Diarization for Long Recordings and Small Collections. *IEEE TASLP* **2012**, 404–413, doi:10.1109/TASL.2011.2162320.
15. Delgado, H.; Anguera, X.; Fredouille, C.; Serrano, J. Fast Single- and Cross-Show Speaker Diarization Using Binary Key Speaker Modeling. *IEEE ACM TASLP* **2015**, 2286–2297, doi:10.1109/TASLP.2015.2479043.
16. Wan, L.; Wang, Q.; Papir, A.; Moreno, I.L. Generalized end-to-end loss for speaker verification. *IEEE ICASSP* **2018**, 4879–4883, doi:10.1109/ICASSP.2018.8462665.
17. Cumani, S.; Brummer, N.; Burget, L.; Laface, P.; Plchot, O.; Vasilakakis, V. Pairwise discriminative speaker verification in the I-vector space. *IEEE TASLP* **2013**, *21*, 1217–1227, doi:10.1109/TASL.2013.2245655.
18. Kenny, P. Bayesian Speaker Verification with Heavy-Tailed Priors. In Proceedings of the Odyssey Speaker and Language Recogntion Workshop, Brno, Czech Republic, 28 June–1 July 2010.
19. Brummer, N.; Silnova, A.; Burget, L.; Stafylakis, T. Gaussian meta-embeddings for efficient scoring of a heavy-tailed PLDA model. *ODYSSEY* **2018**, 349–356, doi:10.21437/odyssey.2018-49.
20. Ramoji, S.; Krishnan, P.; Ganapathy, S. NPLDA: A Deep Neural PLDA Model for Speaker Verification. *ODYSSEY* **2020**, 202–209, doi:10.21437/odyssey.2020-29.
21. Schwarz, G. Estimating the Dimension of a Model. *Ann. Stat.* **1978**, *6*, 461–464, doi:10.1214/aos/1176344136.
22. Chen, S.S.; Gopalakrishnan, P. Speaker, Environment and Channel Change Detection and Clustering Via the Bayesian Information Criterion. In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, VA, USA, 8–11 February 1998; Volume 6, pp. 127–132.
23. Li, R.; Schultz, T.; Jin, Q. Improving speaker segmentation via speaker identification and text segmentation. In Proceedings of the Tenth Annual Conference of the International Speech Communication Association, Brighton, UK, 6–10 September 2009; pp. 904–907.
24. Gupta, V. Speaker change point detection using deep neural nets. *IEEE ICASSP* **2015**, 4420–4424, doi:10.1109/ICASSP.2015.7178806.
25. Siegler, M.A.; Jain, U.; Raj, B.; Stern, R.M. Automatic Segmentation, Classification and Clustering of Broadcast News Audio. In Proceedings of the DARPA Speech Recognition Workshop, Chantilly, VA, USA, 2–5 February 1997; pp. 97–99.
26. Reynolds, D.A.; Torres-Carrasquillo, P. Approaches and Applications of Audio Diarization. *IEEE ICASSP* **2005**, *V*, 953–956, doi:10.1109/ICASSP.2005.1416463.
27. Fukunaga, K.; Hostetler, L. The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition. *IEEE Trans. Inf. Theory* **1975**, *21*, 32–40, doi:10.1109/TIT.1975.1055330.
28. Comaniciu, D.; Meer, P. Mean Shift: A Robust Approach Toward Feature Space Analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 603–619, doi:10.1109/34.1000236.
29. Senoussaoui, M.; Kenny, P.; Stafylakis, T.; Dumouchel, P. A study of the Cosine Distance-Based Mean Shift for Telephone Speech Diarization. *IEEE TASLP* **2014**, *22*, 217–227, doi:10.1109/TASLP.2013.2285474.
30. Vaquero, C.; Ortega, A.; Miguel, A.; Lleida, E. Quality Assessment of Speaker Diarization for Speaker Characterization. *IEEE TASLP* **2013**, *21*, 816–827.
31. Macqueen, J. Some methods for classification and analysis of multivariate observations. In Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 1 Janurary 1967; Volume 1, pp. 281–297.
32. Valente, F.; Motlicek, P.; Vijayasenan, D. Variational Bayesian Speaker Diarization of Meeting Recordings. *IEEE ICASSP* **2010**, 4954–4957, doi:10.1109/ICASSP.2010.5495087.
33. Diez, M.; Burget, L.; Matejka, P. Speaker Diarization based on Bayesian HMM with Eigenvoice Priors. *ODYSSEY* **2018**, 147–154, doi:10.21437/odyssey.2018-21.
34. Villalba, J.; Ortega, A.; Miguel, A.; Lleida, E. Variational Bayesian PLDA for Speaker Diarization in the MGB Challenge. *IEEE ASRU* **2015**, 667–674, doi:10.1109/ASRU.2015.7404860.
35. Viñals, I.; Ortega, A.; Villalba, J.; Miguel, A.; Lleida, E. Domain Adaptation of PLDA models in Broadcast Diarization by means of Unsupervised Speaker Clustering. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017; pp. 2829–2833, doi:10.21437/Interspeech.2017-84.
36. Viñals, I.; Gimeno, P.; Ortega, A.; Miguel, A.; Lleida, E. ViVoLAB Speaker Diarization System for the DIHARD 2019 Challenge. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 988–992.
37. Viñals, I.; Ortega, A.; Villalba, J.; Miguel, A.; Lleida, E. Unsupervised adaptation of PLDA models for broadcast diarization. *EURASIP JASM* **2019**, *2019*, doi:10.1186/s13636-019-0167-7.
38. Diez, M.; Burget, L.; Wang, S.; Rohdin, J.; Cernocký, H. Bayesian HMM based x-vector clustering for Speaker Diarization. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 346–350.
39. Ferràs, M.; Masneri, S.; Schreer, O.; Bourlard, H. Diarizing Large Corpora Using Multi-Modal Speaker Linking. In Proceedings of the Interspeech 2014, Singapore, 14–18 September 2014; pp. 602–606.
40. Nagrani, A.; Chung, J.S.; Zisserman, A. VoxCeleb: A large-scale speaker identification dataset. *arXiv* **2017**, arXiv:1706.08612.
41. Chung, J.S.; Nagrani, A.; Zisserman, A. Voxceleb2: Deep Speaker Recognition. *arXiv* **2018**, arXiv:1806.05622.

42. Bell, P.; Gales, M.J.F.; Hain, T.; Kilgour, J.; Lanchantin, P.; Liu, X.; McParland, A.; Renals, S.; Saz, O.; Wester, M.; Woodland, P.C. The MGB Challenge: Evaluating Multi-Genre Broadcast Media Recognition. *IEEE ASRU* **2015**, 687–693, doi:10.1017/CBO9781107415324.004.
43. Davis, S.B.; Mermelstein, P. Comparison of Parametric Representations for. *IEEE Trans. Acoust. Speech Signal Process.* **1980**, *28*, 357–366.
44. Gimeno, P.; Ribas, D.; Ortega, A.; Miguel, A.; Lleida, E. Convolutional Recurrent Neural Networks for Speech Activity Detection in Naturalistic Audio from Apollo Missions. *Iberspeech* **2021**, 26–30, doi:10.21437/iberspeech.2021-6.
45. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1–32, doi:10.1162/neco.1997.9.8.1735.
46. Villalba, J.; Chen, N.; Snyder, D.; Garcia-Romero, D.; McCree, A.; Sell, G.; Borgstrom, J.; Richardson, F.; Shon, S.; Grondin, F.; et al. State-of-the-art Speaker Recognition for Telephone and Video Speech: the JHU-MIT Submission for NIST SRE18. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 1488–1492.
47. Waibel, A.; Hanazawa, T.; Hinton, G.E.; Shikano, K.; Lang, K.J. Phoneme recognition using time-warping neural networks. *IEEE Trans. Acoust. Speech Signal Process.* **1989**, *37*, 328–339, doi:10.1250/ast.13.395.
48. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 5999–6009.
49. Garcia-Romero, D.; Espy-Wilson, C.Y. Analysis of I-vector Length Normalization in Speaker Recognition Systems. In Proceedings of the Interspeech 2011, Florence, Italy, 27–31 August 2011; pp. 249–252.
50. Villalba, J.; Lleida, E. Unsupervised Adaptation of PLDA By Using Variational Bayes Methods. *IEEE ICASSP* **2014**, 744–748. doi:10.1109/ICASSP.2014.6853695
51. Zhang, A.; Wang, Q.; Zhu, Z.; Paisley, J.; Wang, C. Fully Supervised Speaker Diarization. *arXiv* **2019**, arXiv:1810.04719v4.
52. Blei, D.M.; Frazier, P.I. Distance Dependent Chinese Restaurant Processes. *J. Mach. Learn. Res. JMLR* **2011**, *12*, 2461–2488.
53. Brummer, N.; de Villiers, E. The Speaker Partitioning Problem. *ODYSSEY* **2010**, 194–201.
54. Jelinek, F.; Anderson, J. Instrumentable Tree Encoding of Information Sources. *IEEE Trans. Inf. Theory* **1971**, *17*, 118–119.
55. Brummer, N.; Strasheim, A. AGNITIO's Speaker Recognition System for EVALITA 2009. In Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence, Reggio Emilia, Italy, 9–12 December 2009.
56. Viñals, I.; Ortega, A.; Miguel, A.; Lleida, E. An Analysis of the Short Utterance Problem for Speaker Characterization. *Appl. Sci.* **2019**, *9*, 3697, doi:10.3390/app9183697.